NASA-TM-112430

IN-61-TM
031688

RNS-88-001

Operating Issues for Multiple Cray-2s

William T.C. Kramer
Manager, High Speed Processors
Computational Service Branch (MS 258-6)
NAS Systems Division
NASA Ames Research Center
Moffett Field, CA 94035
(415) 694-4418

Abstract:
The addition of the second UNICOS Cray-2 at NAS has presented some unique technical challenges. There were many issues to deal with in the management of the two systems, ensuring fair and equal treatment of all the users. With the inclusion of the second Cray onto the network, it was possible to tune the systems to provide maximum transfer between the systems. There are also some unique development opportunities now available since both systems have HSX interfaces. Finally, the second system provides the possibility to tune the workload on each system in order to maximize effective use.

This paper will discuss the changes made to the systems because there now are multiple Cray-2s, the management issues involved, the opportunities the multiple systems present, and the on-going development activities at NAS.

Introduction

In April of 1987, the Numerical Aerodynamic Simulation (NAS) Systems Division (NSD) of NASA Ames Research Center agreed with Cray Research to upgraded their first Cray-2, (HSP-1) serial number 2002, from 120 nanosecond memory to 80 nanosecond memory. Instead of an on-site field upgrade, the entire CPU and memory were to be swapped with a new machine — serial number 2013.

In December of the same year, as the first step towards the Extended Operating Configuration[1], NSD signed a contract with CRI to provide the next generation of supercomputer for NAS. In order to provide processing capacity for NAS between December and the possible delivery of the next system (HSP-2) the original Cray-2 (SN 2002) was to remain at NAS at least throughout the year[2]. NAS thus became the first site to have two full size Cray-2 working as partners in the same local network.

---

[1] *Numerical Aerodynamic Simulation Program NPSN System Specification for the Extended Operating Configuration (EOC).* PS-1101-01-C00, NAS Systems Division, NASA Ames Research Center, Moffett Field, CA 94035, September 3, 1986.

[2] Blaylock, Bruce T. and Bailey, F. Ron, *Status and Future Developments of the NAS Processing System Network.* **Third International Conference on Supercomputing**, Boston, Massachusetts, May 15-20 1988.

This paper will describe the configuration decisions made, make comments on the installation, discuss changes needed in the systems when running multiple UNICOS systems and make some observations based on using two UNICOS supercomputers.

Configuring both systems

The HSP-1 system being upgraded had the configuration shown in Table 1. The upgrade also included a High Speed eXternal (HSX) Channel which replaced one of the disk channels.

It has been shown that, for a particular discipline at NAS, the ratio of processing speed to I/O remains constant[1]. What remained was to estimate the throughput improvement of 2013 and adjust the disk storage across both systems as necessary. The performance improvement was estimated by running a suite of benchmark codes first run on 2011 in standalone mode and later on 2013.

The improved CPU time for the 12 codes ranged from 1% to 21%, and the improved throughput from 4% to 26%[2]. These codes are believed to accurately estimate the majority of the work within the NAS Processing System Network (NPSN), and were good indicators of the expected improvement.

The average improvement indicated 2013 processed jobs 15% faster than 2002. Thus, the configuration for 2013 needed 15% more user accessible disk storage than 2002. In addition, storage for system development, program support and source files were to be on 2013. Table 2 shows the general configurations of the systems.

NAS users prefer to have short term scratch disk and permanent disk for their work. Although all our users desire more space, the ratio of 2 to 1 for physical storage has shown to be reasonable. Using the disk quota system developed at NAS, the physical disk is over allocated by a factor of 5 for the scratch file system while the permanent disk is not over allocated. Table 3 shows the distribution of storage for the two systems.

---

[1]Levin, E., Eaton, C.K. and Young, Bruce, *Scaling of Data Communications for an Advanced Supercomputer Network.* **IFIP/IEEE/ITC Third International Conference on Data Communication Systems and their Performance**, Rio de Janeiro, Brazil, June 22-25, 1987.

[2]Smickley, Ronald D., *Acceptance Tests and Performance Measurements Using the NAS HSP-2 Benchmark Programs During the Acceptance Test Periods of the NAS Cray-2 Systems Navier (#2013) and Stokes (#2002)*, **Technical Note No. 4, 88-2000-949**, Sterling Federal Systems, Inc. 1121 San Antonio Ave, Palo Alto, CA 94303, February 29, 1988.

Component	SN 2002
CPUs	4
Memory	256MW 120 Nanosecond DRAM
DD-49s	34
Hyperchannel Adaptors	4
VME Ethernet	1
CTC/3480	1

The original configuration of before the upgrade/installation

Table 1

Component	SN 2002	SN 2013
CPUs	4	4
Memory	256MW 120 Nanosecond DRAM	256MW 80 Nanosecond DRAM
DD-49s	24	31
Hyperchannel Adaptors	4	4
VME Ethernet	1	1
CTC/3480	1	1
HSX	1	1

The basic configuration of the two systems

Table 2

Storage	SN 2002	SN 2013
System Use	7.2	8.4
Development	1.2	4.8
User permanent	7.2	8.4
User Scratch	13.2	15.6
Total	28.8	37.2

Disk distribution for user and system file space

Table 3

Permanent user file storage is separated onto multiple filesystems, most of which use one DD-49, to minimize the impact of disk failure. All these filesystems use the first letter of the host and are distinguished with the second letter being a consecutive letter of the alphabet. These filesystems are mounted under one mount point */u* for user data. The two

systems are names Navier and Stokes[1]. Thus the user permanent file systems on Navier are */u/na*, */u/nb*, */u/nc*, etc. and on Stokes */u/sa*, */u/sb*, */u/sc*, etc. The scratch filesystems consisting of multiple DD-49's are */scr1* and */scr2* on Navier and */scr3* and */scr4* on Stokes.

System Usage

Once the total amount of disk storage was decided (all disk and CPU allocations are assigned once a year on a project basis) it had to be determined which project should be assigned to which system. The increased user disk arriving with HSP-2 allowed disk quota allocations to be increased across the board by 40%. The ratio of user storage space between Navier and Stokes was 55% to 45%.

One of the two options open was to assign projects representing 55% of the total disk space to Navier and 45% to Stokes. The alternative was to assign all the projects, and therefore all the users, to both systems but put 45% of their project quota on Stokes and 55% on Navier.

The first alternative provided less initial administrative work and appeared to be a symmetric solution for using the expected Network File System (NFS). It had the draw back of slow and inaccurate load balancing since it would be done by system administrators based on the past use of the machine. Sharing files between systems would also be difficult, at least until NFS is available and it will be more difficult and disruptive to move users if that became necessary for load balancing. The concern that some users would be placed on the "slower" system was also discussed

[1] Named after the 17th century mathematicians Louis Navier and William Stokes.

The second alternative had the advantages of allowing users to make the choice of the best system for their needs. In most cases (unless development work on items such as HSX connections requires both systems) one of the two systems will be available so users would not have to deal with the supercomputer resource being down. In addition, the user community can load balance the systems themselves by scanning the job queues and determining which is the best system to use. The disadvantages of the second alternative were the potential for duplicate file copies on both systems, the complexity for our users and having the disk storage resources for a project split.

The decision was made to provide accounts for all users and projects on both systems. CPU time would be adjusted for 2002 by the same 15% as our performance studies showed. The goal was to have the same amount of user work completed for the same amount of resource (in this case a CPU hour allocations). Since 2013 was used as a baseline machine, and 2002 CPU hours adjusted downward, the net result was that all NAS users received 18% more processing capacity on the average. Of course, the performance of individual programs varies and some may do better on one system. However, overall, this ratio appears to be accurate.

The files for all projects were distributed on one machine or the other. The system on which a project's files were initially place was termed the "home" system for that project and the other the "remote". Since some projects were using their full allocation they would not fit within the 45/55% split. Hence, a allocation ratio of 60% on the home and 40% on the remote was implemented for each project. Principal investigators could request the allocations for their projects to be changed to any ratio

including 100% on one system as long as the overall ratio of the systems remains approximately 45/55. All software and resources are available on either system.

Figure 1 shows the final configurations of the two systems.

Installation

Installation of 2013 in place of 2002 and the installation of all the new peripherals for 2002 began on January 4, 1988. Prior to the installation, all disks were backed up using 3480 tape drives.

First, the workload tests verified the average 15% throughput difference on the two systems, thereby validating the filesystem distribution. Both systems met or exceeded their contract performance goals.

Some issues did arise during installation which only became apparent because there were two Cray-2's. While the CPU performance tests were standard, the network tests were designed for loop back mode since it was not assumed there would be two system of comparable speed in the network. These tests were adapted to run between the two systems instead of in loop back so no system software was changed. The tests were to have two ftp file transfers over ethernet, two ftp transfers over hyperchannel and a large amount of disk I/O all proceeding simultaneously[1].

There were some problems with the network tests since the buffer size of 4096 bytes did not provide the required hyperchannel transfer rate. On the other hand, a buffer size of 16,777,216 bytes did not allow the ethernet transfers to complete. The buffer size definitions in *if_hy.c* were changed to be specific for the type of network.

Another problem stemmed from the fact UNICOS did not distinguish multiple types of networks. UNICOS tried to acknowledge all received packets through the first device defined in *ioconf.c*.

This problem was solved by making the hyperchannel networks separate special networks and the ethernet the first device in *ioconf.c*. While this solution worked for the performance tests, it was not an acceptable solution for the real network. Separate classes of networks can not be arbitrarily assigned when the systems are attached to nation wide networks. The proper solution will come when UNICOS implements true subnetting.

Once the testing was complete, the reinstallation of user accounts took place. A separate data base was maintained on a support processor with all the project data. Scripts and programs were used to distribute the users across the systems. The distribution was not completely accurate since there was no good method to determine when a file would change from block I/O to track I/O. Indeed, it was estimated that significant disk storage savings could be achieved if the trigger level from block to track I/O were settable on a filesystem basis.

Many other details were handled, such as user access to accounting data on either system, report changes, NQS queue structure changes to allow submission from either system and for completely symmetrical queues. Each system has a script which monitors system usage and manages the number

---

[1] *Initial and High Speed Processor 2 (HSP 2 Computer Systems RFP2-32948(RCB)*, NAS Systems Division, NASA Ames Research Center, Moffett Field, CA 94035, March 4, 1988.

of jobs and types of queues which are running.

Observations

The first observation is that sub-netting is critical. This feature would have solved the problems mentioned above. Now that multiple Crays are in a TCP/IP environment, effective use of the network requires subnetting.

The user load balancing between the two systems appears to work. Both systems are totally busy with average weekly idle time less than 1%. Major changes in usage are generally detected with 12 hours by users, which is faster than operations personnel can respond by moving projects. This quick user response may be due the fact many of our users are familiar with the work of the others and keep up to date on project cycles. Thus, they are able to judge changes in usage quickly and adapt their behavior.

Another need that became obvious was the need for a multi-disk quota system. Disk quota allocations are assigned on a file system basis and balancing projects across these files systems is slow and time consuming. Assigning permanent and scratch quota across multiple file systems provides much greater flexibility.

NFS should be a great help in running multiple systems, particularly when used over an HSX. This is exactly what is planned once version 4.0 is delivered to NSD.

Load balancing between the two systems is feasible and it is possible to tune one system differently than the other. While the ability to dynamically load balance across the two systems is not something that is near at hand, the availability of NFS may make that feasible. The ability to configure the systems for different performance criteria is available and relatively simple.

Conclusion

Having two Cray-2's in the network provide unique challenges and opportunities. The management of the two systems is relatively straightforward and the features of UNICOS allow flexibility in the systems. Networking function can be stressed when two machines were on the system, but these were resolved quickly. Overall, the fact it took less than two weeks to install one system, upgrade another and to adjust the configuration shows the adaptability of UNICOS running in a modular network.

Acknowledgements

The work of making two systems function was shared by many people at NAS. Major contributions were made by the CRI staff, both on site and in Mendota, systems administrators and support staff from General Electric and Sterling Federal Systems, and from other members of the NAS Systems Division.